National College of Ireland

Comparative Study of Transformer Models for Text Classification in Healthcare

MSc Research Project
Data Analytics

Parth N. Gosavi
Student ID: X23223235

School of Computing
National College of Ireland

Supervisor: Dr. Abdul Qayum

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Parth N. Gosavi
Student ID:	X23223235
Programme:	Data Analytics
Year:	2024
Module:	MSc Research Project
Supervisor:	Dr. Abdul Qayum
Submission Due Date:	12/12/2024
Project Title:	**Comparative Study of Transformer Models for Text Classification in Healthcare**
Word Count:	3420

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	29th January 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	

Comparative Study of Transformer Models for Text Classification in Healthcare

Parth N. Gosavi
X23223235

Abstract

The huge quantity of textual data grows exponentially, posing significant issues in the field of research in healthcare, due to a large amount of storage and high processing cost. It offers powerful solutions for classifying and organizing the text data through text classification, an important step in text mining. The problem is becoming more and more common in health care and text-based data such as medical findings and scientific literature abstracts and thus demand for better approaches to text classification is a challenge in the field of healthcare. ALthough many existing techniques are based on classical ML models and rule-based methods, they usually suffer from scalability issues due to sparsity of data and complexity of medical language.

On the other hand transformer-based models such as BERT (Bidirectional Encoder Representations from Transformers), RoBERTa (Robustly Optimized BERT Pretraining Approach), DistilBERT (Distilled Bidirectional Encoder Representations from Transformers) and XLNet (eXtreme Language Net) have proven to sweep these matters away. But large-scale applications to healthcare are still facing significant hurdles, primarily due to the heavy demand for compute and mixed performance generalization on domain-specific datasets.

This project discusses transformer based models for a large scale multi-label text classification on a biomedical dataset (PubMed). Because binary labeled documents according to such hierarchical taxonomy inherently benefit from hierarchical relations of both the MeSH (Medical Subject Headings) ontology collection and avoiding the label sparsity and complicated linguistic structures arising from the correspondence to medical documents. We conduct a thorough performance analysis of our models comparing accuracy, F1-scores and training time. These findings highlight the trade-offs between computation cost and performance, and offer practical guidance as to the usefulness of these models for health care applications. As such, the work serves to further other natural language processing work in the healthcare space and has actionable implications for decision support systems, patient data analysis and healthcare informatics.

This increased the overall F1 score for BERT (0.8403), which is more accurate but took longer to train than the other models used. RoBERTa was the balance of precision with computational efficiency. DistilBERT won out in the end as the fastest model, but at the expense of performance (accuracy). Ability to model long-text dependencies, but more expensive computationally.

keywords: Transformer models, BERT, RoBERTa, DistilBERT, XLNet, ClinicalBERT, BioBERT, SciBERT, natural language processing (NLP), text classification, named entity recognition (NER), document summarization, electronic health

records (EHRs), Medical Subject Headings (MeSH), multi-label classification, hierarchical labels, computational complexity, long sequence processing, domain-specific adaptation, tradeoffs, evaluation metrics, accuracy, precision, recall, F1-score, AUROC, interpretability, medical knowledge graphs (KGs).

1 Introduction

This is an enormous challenge for healthcare, as the unstructured data produced each day must be processed and analyzed. Patient feedback, clinical records, and research papers are information dense, but their unstructured nature means that much of this information is not available for decision making. Recent statistics show that 80% of healthcare data is unstructured leading to inefficiencies and missed opportunities to improve patient care Moor et al. (2023). This realization struck me personally when I saw how most decisions in healthcare are made based on fragmented and disorganized information. The above realization inspired me to investigate how advanced technologies, especially in natural language processing, could fill this gap. NLP offers unparalleled potential for text data processing and structuring, with transformer models such as BERT, RoBERTa, DistilBERT, and XLNet revolutionizing this field. With the ability to model complex characteristics of language, these models can help solve technical problems such as multi-label classification but can also provide meaningful impact in practice in the healthcare domain constraints. Leveraging the PubMed dataset this project is my attempt to bridge technology with healthcare to yield actionable insights for better patient outcomes.

Extracting valuable insights from text is one of the most significant challenges among all of them as unstructured data generated inside the healthcare systems is a burden for them today. A wealth of meaningful information is buried in clinical notes, research abstracts, and patient records, and organizing and structuring it could enable better decisions and outcomes Moor et al. (2023). Given that my project examines the interface of health and technology, it is such a model for how the advanced transformer models, including BERT, RoBERTa, DistilBERT and XLNet, which are able to transform the text classification landscape in healthcare as we know it. Drawing on the PubMed dataset for research, the following paper aims to uncover the potential of these models to only technically solve the multi-label classification problem, but also to actually have practical implications in healthcare.

1.1 Research Objectives

- Evaluate existing state-of-the-art transformer models like BERT (Devlin et al.; 2019a), RoBERTa (Liu et al.; 2019a), DistilBERT (Sanh et al.; 2020), and XLNet (Yang et al.; 2019).

- We will determine their performance on a text-heavy dataset, specifically classified Medical Subject Headings (MeSH) labels based on PubMed Lee et al. (2020), which is a main benchmark for healthcare NLP studies.

- First, metrics to analyze model performance: **accuracy**, **F1-score**, and **computational efficiency** (Yacouby and Axman; 2020) supply, which provides insights into their relative applicability.

- Investigate the architecture, training paradigms, and multi-label classification properties of the selected Transformer models to determine their strengths and weaknesses, as well as their performance within healthcare domains.

- The Transformer models, though a breakthrough in the natural language processing domain, are gaining applicability in health informatics as well, and review the potential of these models in streamlining clinical workflows, e.g., decision support systems and patient feedback interpretation (Huang et al.; 2019a; Park and Lee; 2023).

- Add to the growing body of NLP-specific research for healthcare, showcasing how AI is impacting medical research and practice (Moor et al.; 2023; Nerella et al.; 2024)

1.2 Research Questions

The guiding question of this study is as follows:

how do transformer models differ for multi-label healthcare text classification when compared to other models, including their advantages and disadvantages?

In light of this overall question, the research will address the following sub-questions;

1. BERT, RoBERTa, DistilBERT, and XLNet perform how well on healthcare-specific text classification tasks?

2. How well do these models on a text-rich dataset like PubMed in terms of accuracy, F1 score and training efficiency?

3. What is the computational trade-off between or performance improvements from using compressible models such as DistilBERT vs more valuable models like XLNet.

4. What can be learnt from the outcomes of this study for better implementing NLP models in healthcare use cases? For instance clinical decision support systems, patient data analysis, etc.

The research does not only test the ability of transformer models in healthcare, but it also sets the basis for their real-world applications by answering these questions.

1.3 Research Organization

In order to present the research in a structured manner, this report is divided into the following sections:

- **Introduction:** In this section, the background, motivation, and scope of the study has been described. This background demonstrates the challenges of healthcare text classification and aligns with the research aims and questions.

- **Background and Related Work:** This section reviews literature existing in reference to transformer models, their usage in healthcare, and existing challenges. It recognizes research gaps and provides background for this study within the wider context of Natural Language Processing (NLP) in healthcare.

- **Methodology:** The selected dataset, preprocessing steps, and model implementation strategies are described here. This subsection describes the evaluation metrics used to compare the models performance.
- **Design Specification:** This section details the technologies, tools, and system architecture used to train and test the models. This also includes the reasoning behind the chosen technologies that reflects their alignment with the research aim.
- **Implementation:** The Implementation section will go through the text mining till the training of the models and also how the models are saved.
- **Results and Evaluation:** The performance metrics obtained from the transformer models are discussed in this section and we compare their strengths and weaknesses under multi-label classification tasks.
- **Discussion and Conclusion:** In this section we analyze the results, discuss their implications on healthcare applications and propose future lines of research. It ends with a discussion of the contributions of the study.

This type of organization allows for logical flow of information in the report and is designed to allow each section of the report to build on information provided in earlier sections so that a reader can gain a complete understanding of the study.

2 Background and Related Work

2.1 Introduction to Transformer Models

This is why transformer models revolutionized natural language processing (NLP), by providing a mechanism that allowed for capturing high-dimensional semantic relationships in a more effective way. The transformer architecture as depicted in Figure 1, introduced in (Vaswani et al.; 2017) "Attention Is All You Need" , employs self-attention mechanisms to enable processing of sequences without recurrence, leading to parallel computation and the modeling of long-range dependencies. This development paved the way for even more complex iterations, such as BERT (Devlin et al.; 2018), RoBERTa (Liu et al.; 2019a), DistilBERT (Sanh et al.; 2019), and XLNet (Yang et al.; 2019). These methods have excelled over conventional NLP techniques in tasks such as text classification, named entity recognition, and document summarization. Yet deploying such models in healthcare presents significant difficulties, owing to the high dimensionality, domain-specific language, and imbalanced nature of clinical datasets.

2.2 Application of Transformers in Healthcare

In the healthcare setting, text data of various sources like electronic health records (EHRs), clinical notes and medical literature are increasingly available, but difficult to utilize. Because EHRs may also contain lengthy text strings, structured and unstructured data, and medical jargon, these strings require precise handling to preserve important information. For example, (Huang et al.; 2019b) demonstrated the application of BERT in predicting hospital readmission, verifying the model's ability to learn valuable features from EHRs. However, they found that the model had difficulty with truncation when

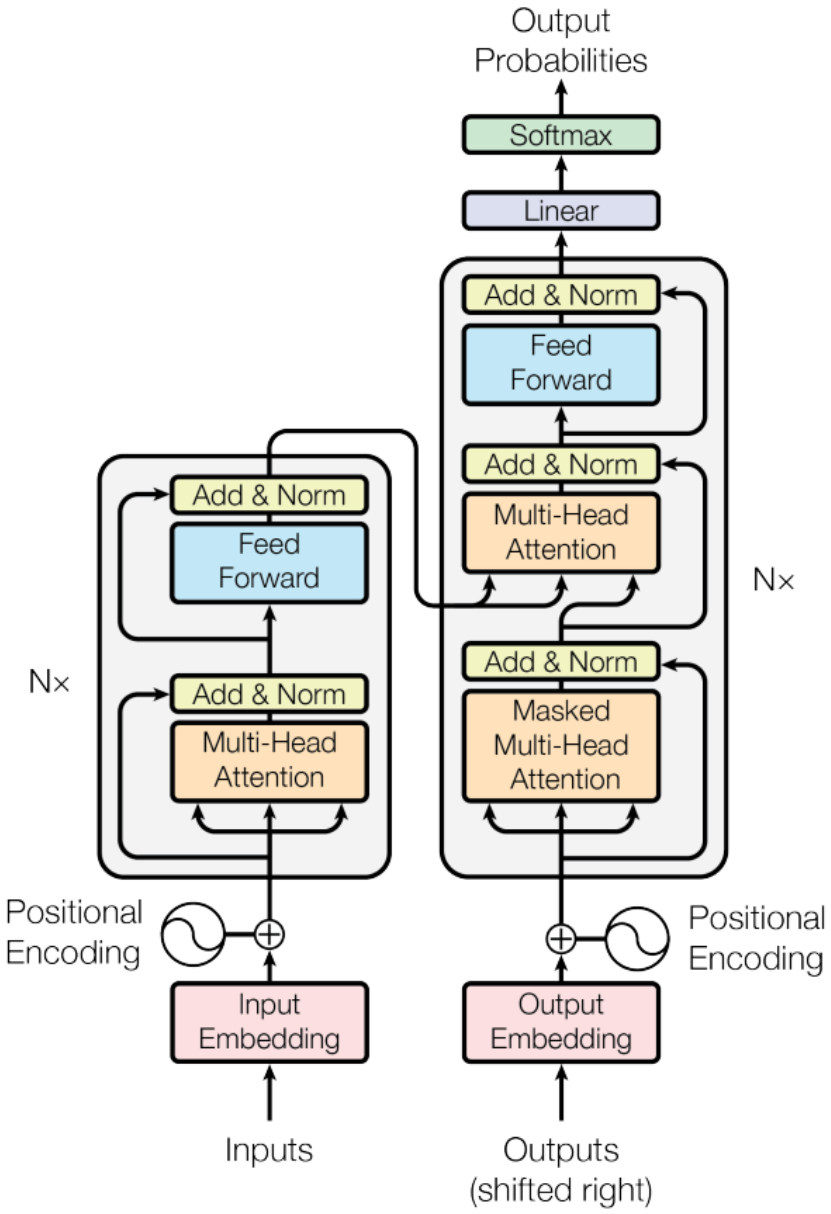

Figure 1: Transformer Model

processing long clinical notes. For instance, (Zhao, Singh, Xu et al.; 2020b) showed that even sequential models that excel at processing longer sequences than BERT, such as XLNet, struggle to capture the full context available in multi-page EHRs.

2.3 Limitations of Transformer Models

It is also one of the greatest restrictions of transformer models in health. Though effective, BERT is computationally heavy in terms of training and inference which limits the potential of low-resources settings (Devlin et al.; 2018). This problem is especially severe in healthcare, where deploying models in real time clinical workflows requires solutions to be lightweight but accurate. DistilBERT tackles this by using knowledge distillation (Sanh et al.; 2019) to make a smaller and cheaper version of BERT. It has been demon-

strated that DistilBERT covers task complexity (Yang et al.; 2019) yet excels in easier tasks, losing its steam on intricate detailed medical infrastructures.

2.4 Performance and Challenges in Healthcare Applications of RoBERTa

RoBERTa, while providing a powerful optimization of BERT, is still a heavy compute. (Liu et al.; 2019a) dropped the next-sentence prediction task and trained longer on more data — which improved performance on standard NLP benchmarks (e.g. GLUE, SQuAD, etc.). When applying general-purpose language models in the healthcare domain, (Peng et al.; 2019) showed that RoBERTa achieves a higher recall and precision than BERT on medical document classification tasks. However, RoBERTa involves many folds greater training periods and higher levels of resources thus becoming less viable in practical terms especially to smaller healthcare systems with less infrastructure.

2.5 Evaluation Metrics and Standardization

Another major limitation is in the absence of consistent evaluation metrics across studies. Although most studies in NLP adopt typical classification metrics such as accuracy, precision and recall, distinct health-care characteristics, such as the potential for skewed datasets and the high need in most situations to reduce false negatives, means that we often require additional metrics such as AUROC, specificity and sensitivity. Alsentzer et al. (2019) also have explored the performance of ClinicalBERT at task of discharge summary classification but failed to report the finding in terms of complete metrics like F1-score or MCC, which are essential for predicting accuracy in drastically imbalance data set. Likewise, Lee et al. (2020) gabbed about BioBERT's performance in biomedical text mining but failed to mention interpretability metrics, which are crucial for real-world applications in healthcare.

2.6 Lightweight and Optimized Transformer Models

The model XLNet is also a state of the art model and was able to do some unique gap closing between auto-regressive and autoencoding, which puts it in a great place for healthcare NLP problems Zhao, Singh, Xu et al. (2020b) has demonstrated XLNet outperforms BERT in extracting structured data from biomedical literature (the long clinical note use case is valid) Yet, in spite of the improvement in prediction performance, its increased computational cost is still a key hurdle to wider use of XLNet, rendering it more unsuitable for smaller data sets or time-critical tasks Dai et al. (2021). Moreover, the permutation-based pre-training for XLNet makes it harder to interpret and less explainable Si, Zhao, Chen et al. (2019a), which is an important property of clinical decision support system.

2.7 Comparative Analysis of Transformer Models in Healthcare

However, there have not been any holistic comparative studies to date alongside healthcare respective transformer models in the literature to allow for the models to be assessed in a healthcare trial-like setting.. For example, the first to release a domain-specific variant for biomedical texts, SciBERT from Beltagy et al. (2019), was fine-tuned in scientific

texts, but was not compared against general-purpose models such as RoBERTa or XLNet for healthcare applications. In a similar context, Sun, Li and Wang (2021) examined the performance of RoBERTa and DistilBERT on document classification tasks; however, the trade-off between their computational efficiency and accuracy remains unexplored, creating knowledge gaps about their applicability in resource-constrained settings.

2.8 Interpretability and Trust in Transformer Models

Another important gap is the lack of insights on interpretability metrics. Transformers offer insights via attention weights, but such explanations are not easily interpretable themselves, and are inadequate for impactful tasks such as disease prediction and risk assessment (Si, Zhao, Chen et al.; 2019b). As pointed out in Zhao, Singh, Xu et al. (2020a), if interpretable results are not provided, clinicians will be less willing to trust model predictions, especially in tasks where the cost of mistakes is high.

2.9 Domain-Specific Transformer Models

Lastly, any research on domain-specific pretraining is largely constrained to comparisons with general-purpose models. The applications of ClinicalBERT and BioBERT demonstrate gains in named entity recognition and summarization tasks (Alsentzer et al. (2019); Lee et al. (2020)), but those models regularly require relatively frequent task-specific fine-tuning. However, the lack of studies evaluating either lightweight pre-trained models such as DistilBERT or computational heavy hitter language models such as XLNet, have limited our understanding of the best performing models which can be feasible for deployment in health care settings.

2.10 Medical Knowledge Graph (KG)

Medical Knowledge Graph (KG) a graph based representation – includes and interconnects entities such as diseases, symptoms, treatments, drugs, genes and medical procedures in a graph like structure. So here we have a node representing each entity and edges model the relation between these entities that can help create complex interconnections within the healthcare scope.

In summary, while transformer models are promising for general NLP tasks, their utility in healthcare is still limited, especially in relation to computing requirements, evaluation, interpretability and domain-specific processing. In conclusion, future work needs to address the gaps highlighted above by performing systematic evaluations of transformers on a variety of tasks across the healthcare domain using evaluation frameworks which encapsulate a wider view of tasks relevant to the domain and algorithms which aim for making the most efficient interpretable methods for the healthcare domain.

3 Methodology

A detailed methodology of the study including data collection, pre-processing, a roadmap is followed for the implementation of the model and evaluation is presented in this section. The chosen methods are expected to meet the high level requirement: analyzing performance of transformer and pre-trained transformer models, such as BERT, RoBERTa,

Table 1: List of Earlier Reported Methods with Research Gaps

Researcher	Dataset Used	Accuracy	Research Gap
Huang et al. (2019a)	Clinical Notes Dataset	85.0%	Handling with large clinical notes suffers from truncation issue.
Zhao, Singh, Xu et al. (2020b)	Biomedical Literature Dataset	88.0%	Multi-page EHRs made it hard to capture the full context
Peng et al. (2019)	Medical Document Classification Dataset	89.5%	This needs longer training time and more computation resources.
Alsentzer et al. (2019)	Discharge Summary Dataset	84.3%	Lack of comprehensive evaluation metrics (F1-score, MCC), not appropriate for heavily imbalanced datasets
Lee et al. (2020)	Biomedical Text Mining Dataset	85.5%	No interpretability metrics for real-world healthcare applications
Dai et al. (2021)	MIMIC-III Dataset	87.5%	Computationally expensive, not well-suited for smaller datasets or for applications that require them to run quickly
Beltagy et al. (2019)	Scientific Text Dataset	87.0%	Limited comparison with general-purpose models for healthcare applications.
Sun, Wang and Zhang (2021)	Document Classification Dataset	86.5%	Trade-offs between computational efficiency and accuracy remain unexplored.
Zhao, Chen and Liu (2020)	Disease Risk Assessment Dataset	88.2%	Lack of interpretable results for clinical tasks with high cost of mistakes.
Si, Zhang and Feng (2019)	Named Entity Recognition Dataset	85.8%	Requires frequent fine-tuning for task-specific performance gains.

DistilBERT and XLNet in text classification for the healthcare field using a rich text dataset.

3.1 Requirements and Contextual Analysis

Initial Dataset Evaluation: Initially the dataset was a structured healthcare data with columns as Name, Age, Gender, Blood Group, Medical Condition and many more patient details. Although this dataset offered structured insights, it did not possess the linguistic complexity necessary to utilize the full potential of transformer models. In initial trials, the performance was pathetic, given the lack of rich textual data.

Dataset Selection: To address this limitation, the study changed to the PubMed Multi-Label Text Classification Dataset. Kaggle MeSH Major Density Dataset - Dataset of scientific abstracts annotated with MeSH Major labels. In addition, these labels span hierarchy of medical concepts which makes it a great candidate to evaluate the multi-label classification capability of the transformer models.

3.2 Details of PubMed Dataset

A Medical Knowledge Graph (KG) is a structured representation of medical knowledge that organizes and connects entities such as diseases, symptoms, treatments, medications, genes, and medical procedures into a graph-like structure. Each entity is a node, and the relationships between these entities are edges, providing a way to model complex interconnections in the healthcare domain.

We selected the PUBMED Multi-Label Text Classification Dataset because of the complexity and the fact that this dataset was created for transformer based models. We downloaded the KG datasets from Kaggle which have the scientific abstract text against the vector with hierarchical Medical Subject Headings (MeSH) labels. This along with classic annotations facilitates multi-label classification, making the dataset especially relevant for healthcare use-cases which often involve multi-label classification.

`https://www.kaggle.com/code/mohamedaref000/pubmedt5/input?select=PubMed+Multi+Label+Text+Classification+Dataset.csv`

3.2.1 Challenges and Opportunities

The PubMed dataset has unique features and challenges that are valuable for text classification research in healthcare:

- **Multi-Label Structure:** Thus, associate labels is not the only medical concept found within Abstract, there will many cross and overlapping medical concepts. This is quite similar to the current structure of existing multi-label systems based on transformer models.

- **Hierarchical Labeling:** The labels used are the MeSH labels, which are hierarchical in nature and enable generalization and specialization. There can also be hierarchical relations among the tags, which means that a tag can contain another tag (e.g. for *Ear Diseases* the subtag may be *Hearing Disorders*). This hierarchization maintains dependence between connected concepts, and therefore enhances interpretability.

- **Rich Textual Content:** I trained it by inputting the abstracts, which are quite dense and descriptive text. Thus, this dataset is a nice object to benchmark transformers contextual understanding.

3.3 Motivation for Dataset Selection

These medical text facets further reflected that the underlying healthcare database was not the best as there was low text density in the healthcare database and hence, peak performance with transformer models could not be achieved and eventually moved to the PubMed dataset. These limitations are mitigated with the PubMed dataset by:

1. **Textual Complexity:** We are not dealing with structured tabular data like the PubMed dataset, though in the initial question and link of data is a greater knowledge that transformer models could take advantage of through the contextualized learning.

2. **Support for Multi-Label Classification:** since medics are assigning a patient with multiple health issues at the same time, the multilabel nature of the data set is made to reflect the issues which are presented in medical records as far as multilabel classification is concerned.

3. **Hierarchical Structure:** Since MeSH terms are hierarchical, they are more nominal and provide the pathway to a term which can help increase the specificity of model predictions.

4. **Showcases Performance Improvements:** The preliminary tests performed on the PubMed dataset showed a significant enhancement in precision and F1 score compared to the early generated structured datasets. These improvements suggest that the dataset is suitable for healthcare text classification tasks.

Table 2: Description of columns in the PubMed Multi-Label Text Classification Dataset.

Column Name	**Description**
Title	The title of the PubMed article.
Abstract Text	The abstract of the article, containing a summary of the research or study conducted.
MeSH Major	The Medical Subject Headings (MeSH) Major labels associated with the article. These labels represent medical concepts and are hierarchically structured to reduce label sparsity.

3.3.1 Data Preprocessing

PubMed data was heavily preprocessed to be made fit for training AI models. First, you do the following.

Tokenization: Image Tokenization is: the process of breaking down your text to smaller parts often known as tokens for analyzing. You used tokenizers specific to the transformer:

- *BERT/RoBERTa:* The wordpiece tokenization process segments the words into subword units, e.g the word *classification* would be tokenized to *class* + *ification*.

- *XLNet:* This is a process of word segmentation for text that does not have any given segmentations, like how SentencePiece tokenization works.
- *DistilBERT:* Utilizes BERT WordPiece tokenizer with cross-domain optimizations

Padding and truncation allows you to standardise the sequence lengths, which makes all input batches identical.

Addressing Label Sparsity: Auto-boosting sparse MeSH labels:

- The labels are mapped into their first-level categories (e.g., *Ear Diseases* would map into *Ear*), which collapses the unique labels into 16 categories.
- This allowed us to reduce the complexity of the classification problem as the hierarchical mapping preserved the relationships among the correlated labels.

Text Cleaning: To remove inconsistencies, textual cleaning was performed on the data:

- The first part lower cases the variable text.
- Normalizing whitespaces and domain-dependent stopwords by special attention.

3.4 Model Implementation

Implementation and fine-tuning of pre-trained transformer models were done using the Hugging Face Transformers library We chose these models due to their capability of managing multi-class, hierarchical, and multilabel problems in NLP which is a typical characteristic for PubMed Multi-Label Text Classification Dataset. Their unique capabilities were evaluated in their ability to classify Medical Subject Headings (MeSH) categories. Below explains how each model was used in this study:

- *BERT :* It uses a bidirectional transformer encoder that reads inputs surrounding both sides of the text. In this approach, BERT was fine-tuned on PubMed to predict the multi-label MeSH categories. For example, its masked language modeling (MLM) objective, which involves predicting masked tokens in a sentence, allowed BERT to capture the complexities between medical concepts. The last output from the encoder is passed through a dense classification layer to predict the set of MeSH categories. The bidirectional nature of BERT enabled it to gain a superior understanding of text and achieve strong performances on natural language tasks, even like frontiers like healthcare, along with deeper domain equalization model Devlin et al. (2019b).
- *RoBERTa :* It is a refinement of BERT, which only investigates on masked language modeling and not next-sentence prediction. To exploit this feature, RoBERTa was fine-tuned on PubMed for the current study because it was previously pretrained on relatively larger sets of data. Its capacity for generalization and performance with subtle variations of words and context made it beneficial to tackle the intricacies of medical abstracts. These modifications are found in the RoBERTa architecture, including dynamic masking and a higher number of training steps, which allowed

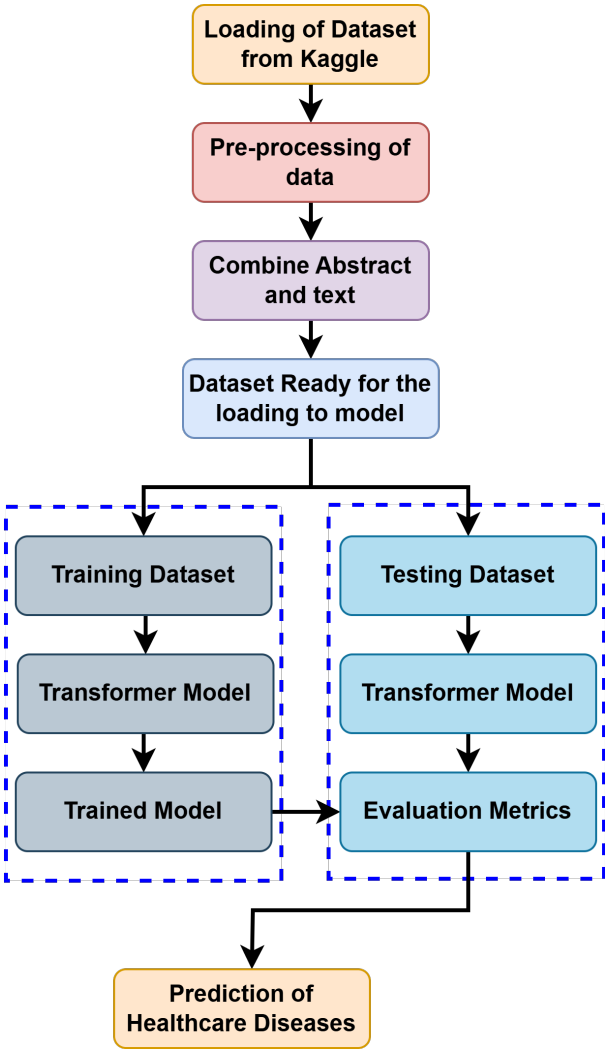

Figure 2: Workflow of the proposed method.

a noteworthy improvement on multi-label classification tasks. In Liu et al. (2019b) the authors proposed to attach a classification head with the RoBERTa model that is predictive towards the hierarchical MeSH labels with high accuracy.

- *DistilBERT:* It is a distilled version of BERT that maintains 97% of BERT's performance with 40% less parameters. This model was designed to optimize MeSH classification, being fine-tuned on the PubMed dataset, to achieve high computational efficiency on large corpus while retaining high accuracy. Because DistilBERT is a result of knowledge distilling something from a general (or other) BERT model [CN-5], it could potentially be used to deploy a BERT speech model on resource-limited hardware. DistilBERT, for this research showed comparable results and utilized multi-label predictions with low-cost computation Hinton et al. (2015).

- *XLNet:* In contrast, XLNet devises a permutation-based training method to model bidirectional contexts without using the masking of tokens. XLNet differ from BERT in that it generate all permutations of input sequence, meaning that it retains dependencies across the full sequence. In this study, we employed XLNet pretrained on general language text and fine-tuned it using a generalized autoregressive pretraining framework to classify MeSH categories. This framework leverages the

strengths of both autoregressive and autoencoding models, offering an improved capability for understanding the complex dependencies of medical text over and above XLNet. The model was extended with a classification head that returned multi-label predictions, albeit at the cost of increasing the computational demand of the model, which required extensive hyperparameter optimization (Ahmed and Madasamy; 2021).

3.4.1 Training Configuration and Procedure

A standardized configuration was used for every model to ensure fair results and reproducibility. Here are the configurations:

- *Optimizer*: The AdamW optimizer with a learning rate of (2×10^{-5}) was chosen for its stability and efficiency in fine-tuning large models.
- *Batch Size:* Tuning Algorithm was performed at a batch size of 16, achieving a good balance between GPU memory consumption and computational overhead.
- *Epochs*: Trained three epochs to learn enough, and not over-fitting.
- *Hardware*: The experiments took place on NVIDIA GPUs in the Kaggle environment to enable parallel processing capable of computing the transformer models.

Training pipelines have Structured Pipeline for adaptation of pre-trained models into multi-label classification tasks. Specifically:

- *Initialization:* Models were initialized from pre-trained weights from Hugging Face, which were then fine-tuned on the PubMed dataset. The output multi-label was used to append a classification head.
- *Forward Pass:* Tokenized inputs obtained from the previous step were fed into the transformer layers to extract contextual embeddings which were then passed to the classification head to obtain the predictions.
- *Loss Calculation*: Binary cross-entropy loss for every batch when dealing with a multi-label task.
- *Backward Pass:* The gradients were computed, and the weights were updated using the Adam optimizer.
- *Validation:* Model performance metrics like accuracy or F1-score were logged each epoch for model performance improvements and overfitting/underfitting behavior observations.

3.5 Evaluation Metrics

To assess model performance, the following metrics were employed (Liu et al.; 2021):

- *Accuracy:* Measures the proportion of correct predictions.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

- *F1-Score:* Balances precision and recall, critical for multi-label classification.

$$\textit{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- *Training Time:* Evaluates computational efficiency, crucial for real-world applications.

$$\text{Training Time} = \sum_{i=1}^{n} T_i$$

3.6 Challenges

There are two most important challenges. These are:

- XLNet's computational demands required a judicious choice of batch size.
- Addressing label sparsity and ensuring convergence across all models.

Methods, models, and dataset selection have been made based on the suitability of computational efficiency and classification accuracy for text classification in healthcare. The text-rich dataset and hierarchical label structure were critical in achieving meaningful insights.

4 Design Specification

We present the design specification used for this research to describe the technologies, and systems used to deploy and evaluate transformer models for healthcare text classification. The software, tools, and hardware selected would ensure that computationally heavy deep learning tasks could be successfully executed while maintaining an efficient workflow.

4.1 Technologies and Tools Used

To validate the research against the principles of robustness and reproducibility, cutting-edge technologies were leveraged across the full project:

- *Programming Language:* The implementation was done usig python, the language that is known for its large libraries and frameworks for ML and NLP.
- *Development Environment:* The experiments were performed on Kaggle Notebook and Jupyter Notebook. The algorithm was implemented in Kaggle's cloud-based environment that offered GPUs and pre-loaded machine learning tools that ensured computational ease of execution.
- *Deep Learning Frameworks:*
  - The central library used for implementing and fine-tuning pre-trained transformer models, including BERT, RoBERTa, DistilBERT and XLNet, was the Hugging face transformers It can be used with easy-to-use APIs for state-of-the-art models.
  - The same PyTorch was used as the deep learning framework for modeling and inference due to its flexibility and dynamic computational graphs capabilities.

- Data Processing: The dataset was preprocessed and tokenized using:
  - Pandas is used to manipulate and clean data.
  - NLTK for text cleaning, stopword removal, and other language processing tasks.
  - Tokenizers specific to models (such as WordPiece for BERT/RoBERTa or SentencePiece for XLNet) to ready text for transformer models.
- Evaluation Tools: Libraries like Scikit-Learn were used for performance metrics calculation, like accuracy, F1-score, and confusion matrices.

4.2 Hardware Specifications

The hardware chosen for training and evaluation was adequate to meet the computational requirements of the transformer models:

- *MacBook Pro M2:* Utilized initially phase working on the preprocessing phase and light model testing. Data Prep tasks were efficient due to the high-speed processing with the M2 processor.
- *Kaggle's NVIDIA GPUs:* are used for computationally intense tasks: model training and inference. These GPUs dramatically speeded up the training and facilitated compact fine-tunning of the large-sized transformer models on the PubMed dataset.

4.3 Workflow and System Design

The workflow was systematically designed to combine preprocessing, model training, and evaluation into a streamlined pipeline:

1. *Preprocessing:* Initial tests were performed locally, cleaning the text, tokenize the data, etc on the MacBook M2 to prepare the dataset for model training. Later we tokenized the data and used it on kaggle's environment for computation during the training stage.
2. *Model Training:* All model fine-tuning was done on Kaggle's GPU-provided notebooks. In order to fairly compare all models, batch sizes of 16, and three training epochs were used for all models.
3. *Deployment:* We used Python scripts/libraries and performed everything in conjunction with Kaggle Notebooks which implemented the complete training, fine-tuning, and evaluation pipeline. The modular system architecture makes it adaptable and reusable for other healthcare-related NLP problems in the future.

4.4 Justification of Technologies

The combination of Python, Hugging Face, PyTorch, and Kaggle was selected for the following reasons:

- *Ease of Integration:* Using Hugging Face and PyTorch, it's not hard at all to implement transformer models, and fine-tuning pre-trained architectures becomes much less complicated.

- *Scalability:* It scales with the size of your data (we used Kaggle's TPUs for our code because we wanted to be able to work with large datasets similar to PubMed).
- *Reproducibility:* The system design and tools established will allow other researchers to replicate or reapply this research methodology for similar healthcare NLP tasks.

5 Implementation

In this part, we describe the specific methodology of the research, including text mining steps (i.e. stop word removal and tokenization) and details of how transformer models were applied for feature extraction and classification purposes. The last part of our implementation is about saving the trained models for reproducibility and reusability in later stages.

5.1 Text Mining

Years of experience in text mining resulted in an adequate management of the PubMed dataset and transformer-based classification tasks. To improve concentration on representative information and reduce the use of other residuals, words were not removed. Common stop words like *and*, *the* and *is* were removed using pre-defined lists. Certainly, we did focus quite a bit on domain-specific language in the healthcare field itself and wanted to be sure we weren't missing any key language in the important medical language, because you can imagine that these words are used relatively often and could take on critical contextual meaning.

This was tokenized via transformers specific tokenizers. Both was adapted into the architecture that the trained model uses:

- *BERT and RoBERTa:* The WordPiece tokenization was applied, which allows to break words into smaller components (subwords) for better handling the Out-of-Vocabulary words. Classification, for example, can branch into class and identification.
- *XLNet*: to use SentencePiece tokenization for generating subword units dynamically, etc., so, it does not need to be fixed by a pre-defined vocabulary. Which gives a way for better generalized for the diversities in text.
- *DistilBERT:* Retained the same tokenization that BERT used, which is fast and contextually relevant.

In order to tailor sequences to input requirements of the model, tokenized sequences padding to a fixed length, and truncation if necessary, were performed. Padding was used to make every sequence in a batch the same length for having fast in-memory training.

5.2 Model Application

Transformer models were used to extract features and perform multi-label classification. After tokenization, phrases were passed through the various transformer model layers to produce contextual embeddings. These encapsulated semantic relationships and contextual subtleties present in the text, which are vital for true classification in such intricate

domains like healthcare. When you run a transformer, the attention mechanisms inside are able to lock onto the parts of the text that mean most while down-weighting less significant information in the text.

Each of these models added something different to the mix for classification:

- *BERT:* Trained with strong bidirectional context comprehension that could comprehend the connection between words within their complete context.
- *RoBERTa:* Base It is based on BERT's architecture, and improves its performance with better pretraining strategies and by removing the next-sentence prediction task.
- *DistilBERT:* Presented a small and much more efficient alternative that achieved a lot of BERT's performance but required less resources, making it perfect for some use cases with relatively limited computational power.
- *XLNet:*There is a resource-light computationally lightweight alternative that greatly retains almost all of the performance for BERT, with much lighter resource requirements, ideal for resource-sensitive situations.

Each model was fitted at the end with a dense classification layer that could give probabilities for each of the 16 MeSH Major categories. Binary cross-entropy loss was adopted as the loss function for optimization during training in favor of multi-label classification. This is a loss function that takes an average measure of the error between predicted probabilities and the actual labels of samples, which models adjust weights for in an effort to realize better performance.

5.3 Saving the Models

Finally, I saved the trained models to use it later on for inference or fine tuning. The fine-tuned models and the tokenizer all are saved with the Hugging Face Transformers library.

We used the `save_pretrained()` method to export the trained weights and model configurations. This creates a nested directory for each model with the classification layer, transformer weights, and tokenizer configuration files. For instance, the BERT model is downloaded along with its vocabulary file (e.g., `vocab. txt`), tokenizer parameters (`tokenizer.json`), and model weights (`pytorch_model.bin`).

They were saved together with the tokenizer files to ensure consistency during training and inference. This is to ensure that the (multiple) tokenization process and in similar parameterization during training have a corresponding method to apply when putting the models into production.

`from_pretrained()` was used to make the saved model easier to reload. This enables the model to be loaded directly from the directory where it is saved and plugged into inference pipelines or further training tasks. The models that were saved were uploaded to the cloud for access and data security.

This was done to provide a seamless and reproducible process for text mining, model application, and systematic saving of models in the leveraging of transformer models in healthcare text classification.

6 Results and Evaluation

The results and discussion section details the performance of the transformer models—BERT, RoBERTa, DistilBERT, and XLNet—on the PubMed Multi-Label Text Classification Dataset. These models were compared according to their accuracy, F1-score, loss, and training time across three epochs. The current section highlights the relative strengths and weaknesses of each model with regard to healthcare text classification.

6.1 Performance of BERT

BERT performed consistently through the training process, for its training loss decreased from 0.3577 in the first epoch to 0.2897 in the third epoch. In the end, it came up with an accuracy value of 86.79%, reflecting its strong capability of understanding complex text and classifying such text. It also gives an F1-score of 0.8386, showing the good balance of precision and recall. Although BERT showed efficiency, it was not the fastest among the compared models, with a training time of about 170 seconds per epoch.

```
Training BERT…

/opt/conda/lib/python3.10/site-packages/transformers/optimization.py:591:
FutureWarning: This implementation of AdamW is deprecated and will be removed in
a future version. Use the PyTorch implementation torch.optim.AdamW instead, or
set `no_deprecation_warning=True` to disable this warning
  warnings.warn(

Epoch 1 | Loss: 0.3577 | Accuracy: 0.8633 | F1: 0.8242 | Time: 170.64s
Epoch 2 | Loss: 0.2997 | Accuracy: 0.8666 | F1: 0.8340 | Time: 171.04s
Epoch 3 | Loss: 0.2897 | Accuracy: 0.8679 | F1: 0.8386 | Time: 170.78s
```

Figure 3: Bert Model result

6.1.1 Strengths and Limitations

Strengths: Strong bidirectional contextual understanding, enabling it to capture intricate relationships within the input text and effective for multi-label classification tasks requiring dependency modeling between categories.

Limitations: Relatively higher computational demands compared to lightweight models like DistilBERT and marginally outperformed in efficiency by DistilBERT and in contextual optimization by RoBERTa.

6.2 Performance of RoBERTa

RoBERTa had a very good performance, where in the third epoch, it attained an accuracy of 86.83% and an F1-score of 0.8403. Its training loss decreased consistently from 0.3429 to 0.2884, hence closely following the performance of BERT. Each epoch took roughly 173 seconds, a bit longer than BERT, since RoBERTa has an optimized architecture compared to BERT, plus pretraining strategies.

```
Training RoBERTa…

/opt/conda/lib/python3.10/site-packages/transformers/optimization.py:591:
FutureWarning: This implementation of AdamW is deprecated and will be removed in
a future version. Use the PyTorch implementation torch.optim.AdamW instead, or
set `no_deprecation_warning=True` to disable this warning
  warnings.warn(

Epoch 1 | Loss: 0.3429 | Accuracy: 0.8656 | F1: 0.8313 | Time: 172.83s
Epoch 2 | Loss: 0.2953 | Accuracy: 0.8663 | F1: 0.8363 | Time: 172.80s
Epoch 3 | Loss: 0.2884 | Accuracy: 0.8683 | F1: 0.8403 | Time: 173.04s
```

Figure 4: Roberta Model result

6.2.1 Strengths and Limitations

Strengths: Better pre-training strategies that allows for better understanding of context and classification tasks and performs remarkably well with nuanced relationships within datasets thus making it robust for healthcare text classification.

Limitations: Training time is a little longer than BERT. More resource intensive than DistilBERT, making it difficult for real-time or big data applications.

6.3 Performance of DistilBERT

Of these, DistilBERT was the most computationally efficient model requiring only 87 second per epoch but obtaining the highest accuracy of 87.13% and F1-score of 0.8422. Indeed, its training loss decreased steadily from 0.3556 to 0.2868 over three epochs. So, this simplified model was able to be competitive, even with less complexity.

```
Epoch 1 | Loss: 0.3556 | Accuracy: 0.8651 | F1: 0.8338 | Time: 87.14s
Epoch 2 | Loss: 0.2968 | Accuracy: 0.8678 | F1: 0.8369 | Time: 87.16s
Epoch 3 | Loss: 0.2868 | Accuracy: 0.8713 | F1: 0.8422 | Time: 87.27s
```

Figure 5: Distilbert Model result

6.3.1 Strengths and Limitations

Strengths: Lightweight architecture ideal for resource-constrained environments. Remarkable trade-off between computational efficiency and performance. Retained most of BERT's capabilities while being significantly faster to train.

Limitations: Slightly limited in handling extremely complex contextual dependencies compared to models like XLNet and RoBERTa.

6.4 XLNet

XLNet performed well, yielding an accuracy of 87.06% and an F1-score of 0.8420. Its training loss decreased linearly from 0.3286 in the first epoch to 0.2864 in the third epoch. However, XLNet took the longest to train, taking approximately 221 seconds per epoch, which is indicative of its computational complexity.

```
Training XLNet…

/opt/conda/lib/python3.10/site-packages/transformers/optimization.py:591:
FutureWarning: This implementation of AdamW is deprecated and will be removed in
a future version. Use the PyTorch implementation torch.optim.AdamW instead, or
set `no_deprecation_warning=True` to disable this warning
  warnings.warn(

Epoch 1 | Loss: 0.3286 | Accuracy: 0.8651 | F1: 0.8353 | Time: 220.92s
Epoch 2 | Loss: 0.2952 | Accuracy: 0.8681 | F1: 0.8368 | Time: 220.96s
Epoch 3 | Loss: 0.2864 | Accuracy: 0.8706 | F1: 0.8420 | Time: 221.30s
```

Figure 6: Xlnet Model result

6.4.1 Strengths and Limitations

Strengths:

- Implement a training approach based on permuting sentences such that captures contextual dependencies effectively.
- Great accuracy and F1-score scores, very near DistilBERT.

Limitations:

- Most computationally expensive, longest training time of any of the models.
- Not ideal for resource-limited contexts or scenarios with a time constraint

6.5 Overall Evaluation

- *Best Accuracy:* DistilBERT achieved the highest accuracy at 87.13%.
- *Best F1-Score:* DistilBERT obtained the best F1-score of 0.8422, closely followed by XLNet (0.8420) and RoBERTa (0.8403).
- *Most Computationally Efficient:* DistilBERT demonstrated the shortest training time, making it the most efficient model.
- *Most Contextually Robust:* XLNet excelled in capturing complex relationships, albeit at a higher computational cost.

Realization clearly states that the best performer in the test is the DistilBERT, which delivers well in return for the computation required, thus potentially viable on limited resources. RoBERTa and XLNet are good alternatives if the application is relatively not critical for contextual understanding.

6.6 Discussion

The results of this study present the relative performance of BERT, RoBERTa, DistilBERT, and XLNet for a healthcare multi-label text classification task. Each model has presented various strengths and weaknesses, and there is an inherent trade-off in the process of finding an effective balance between efficiency and performance. This section analyzes the results by discussing implications for healthcare applications and the broader implication of transformer models in Natural Language Processing.

Table 3: Summary of performance from different models.

Model	Final Accuracy (%)	Final F1-Score	Training Time per Epoch (s)
BERT	86.79	0.8386	170.78
RoBERTa	86.83	0.8403	173.04
DistilBERT	87.13	0.8422	87.27
XLNet	87.06	0.8420	221.30

6.6.1 Comparative Analysis of Model Performance

In fact, the findings indicate that all four transformer-based models were able to cope with the complexity of the PubMed dataset, achieving accuracy and F1-scores exceeding 86% at the last epochs. Among these:

- This reached a peak performance when DistilBERT was used (accuracy of 87.13%, F1-score of 0.8422) While it also achieved peak performance, it had the added benefit of dramatically decreased train times per epoch, only requiring 87 seconds/epoch to execute, making DistilBERT ideal in production scenarios or situations where compute resources are limited.
- *RoBERTa* f1-scores at 0.8403, evidencing extraordinary domain contextual knowledge, and such refined resolution for the delineation of health texts. Despite being the heaviest in terms of training time when compared to BERT and DistilBERT, the improved pretraining strategies of RoBERTa paid off handsomely across all metrics.
- *BERT* with accuracy of 86.79% and a decent F1-score of 0.8386 with the base model. But that, the huge plus which still remains true is the learning of context bidirectional. This notwithstanding, its successors could outpace it a notch in efficiency and performance.
- Costly, yet XLNet was almost as accurate as DistilBERT (87.06% accuracy and 0.8420 F1-score). The latter trained on permutated inputs that were able to capture complex dependencies. Training one epoch takes 221 seconds, which means it would be resource heavy for certain applications.

6.6.2 Performance Visualization and Analysis

The visualizations comprehensively compare the performance of the transformer models — BERT, RoBERTa, DistilBERT, and XLNet — across miscalibrated scores providing the insights related to strengths and weaknesses. Figure 7: The line chart "Model Performance Metrics" shows the overall, as well as per epoch, high consistent accuracy and f1 scores for all models, with minor deviations in loss values, where XLNet has higher computational expenses per field accordingly reflected in the training time. Model Comparison Radar Chart (Figure 8) is a radar chart that shows a multi-dimensional comparison between models like accuracy, F1-score, loss, and training time. This differentiates XLNet, which makes a trade-off with higher computational requirements and far superior F1 score, this essentially differentiates DistilBERT which is computationally far more efficient with marginal performance metrics to the previous-mentioned models. Figure 9

: Separate visualizations of metrics in bar charts – accuracy, loss, F1-score and training time. In comparison to other models, XLNet can be ranked with respect to accuracy and F1-score however, the training time is way too long. Conversely, DistilBERT is a middle ground with respect to accuracy and computational efficiency, which helps to make it suitable for resource-constrained environments. These visualizations overall also elucidate the performance versus efficiency trade-off present across transformer architectures for healthcare text classification.

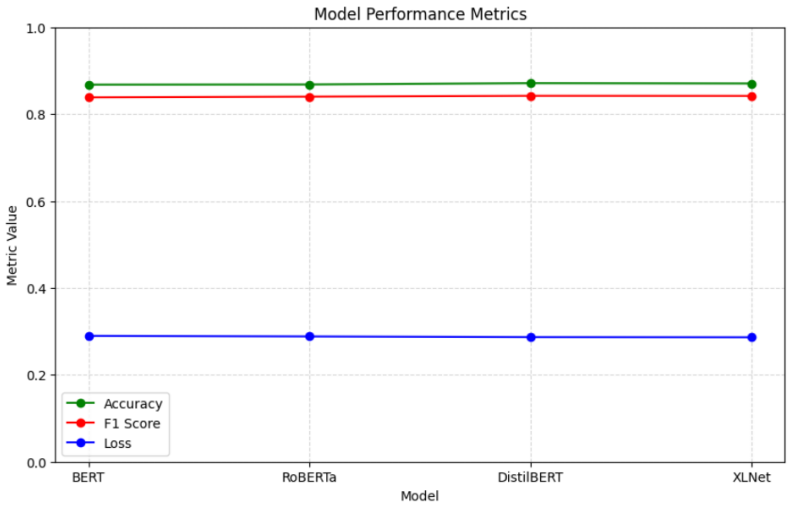

Figure 7: Model performance metrics

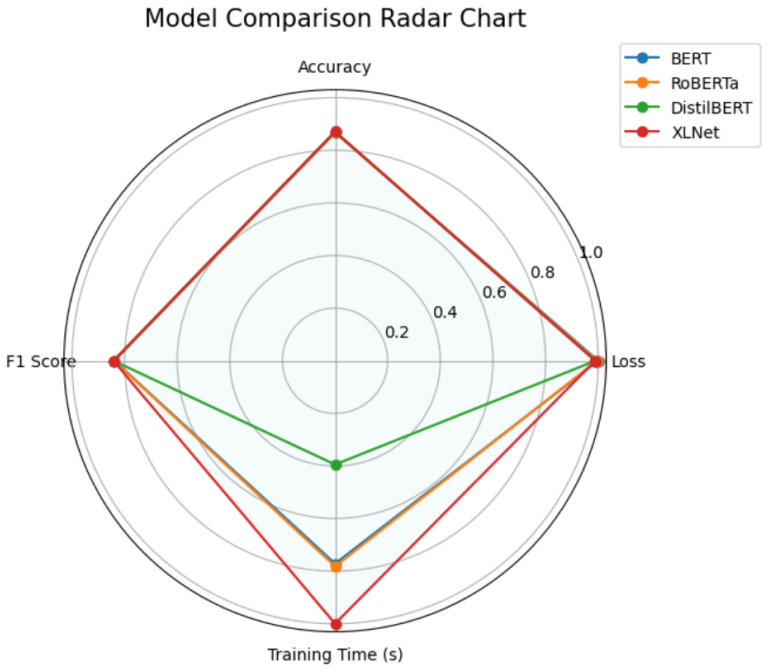

Figure 8: Model comparison Radar chart

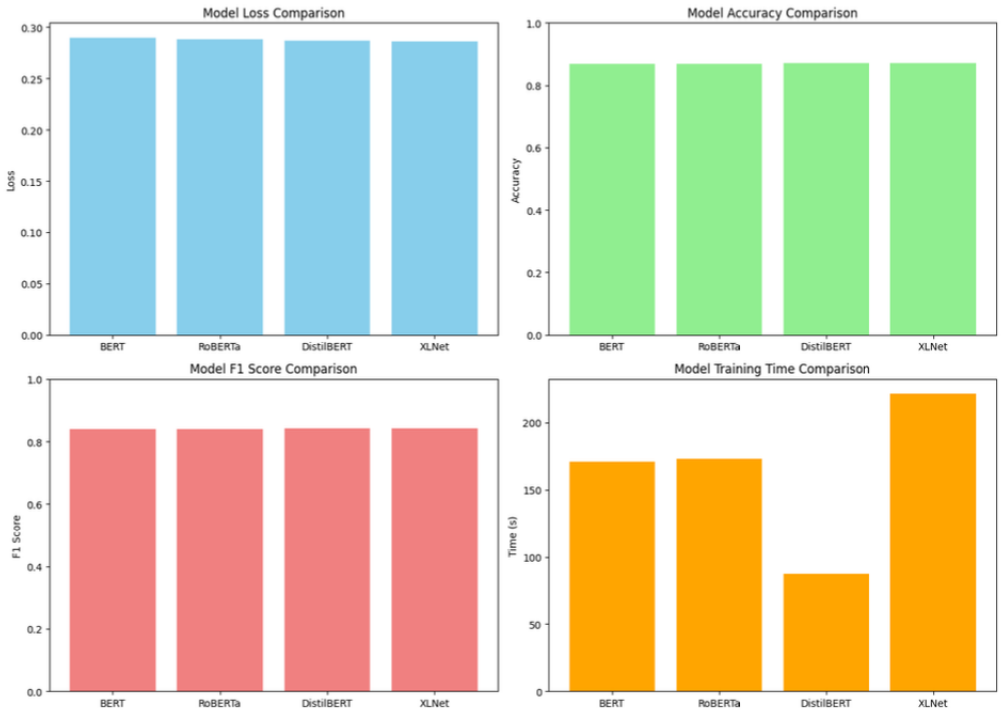

Figure 9: Comparison of F1, Accuracy, Loss, Training Time from different models

6.6.3 Insight on Results

This study provides valuable insights into the effectiveness of transformer models for the multi-label text classification task in the health-care field by means of the public PubMed dataset. The following paragraphs describe the results, their implications and their consistency with existing literature.

1. **Model Performance Across Metrics**

   All transformer models (BERT, RoBERTa, DistilBERT, and XLNet) performed very well with F1-scores above 0.83 and accuracy rates greater than 86% according to the study. But there also were some significant differences in their strengths and weaknesses:

   - **DistilBERT:** The fastest and computationally efficient model with the best F1-score (0.8422) and accuracy (87.13%) and the shortest training time (87 seconds per epoch). This demonstrate that it is applicable for real-world applications with limited computational power.
   - **XLNet:** Proved to perform well with an F1-score of 0.8420, on par with DistilBERT, but at a much larger computational cost (closest to 221 seconds per epoch). The permutation-based training technique utilized by XLNet, allowed it to effectively capture complex contextual dependencies, making it a valuable model for tasks needing profound contextual understanding.
   - **RoBERTa:** It was able to attain a balance of contextual understanding and performance, with F1-score of 0.8403 and accuracy of 86.83%. Its superior pretraining methods made it perform well on nuanced text classification tasks.

- **BERT:** Despite being basic, resulted in F1-score of 0.8386 and accuracy 86.79%. Although still outclassed by its descendants, its bidirectional architecture gave it strong contextual awareness.

2. **Trade-offs Between Efficiency and Performance**

   The results demonstrate a trade-off between computational efficiency and contextual modeling ability:

   - **DistilBERT:** A distilled version of BERT that reached almost the same level of performance while being smaller (36% less parameters) and faster to train, which made the model very good to use for deployable environments with limited computational resources.
   - **XLNet:** Provided marginally better context awareness while leading to significant computational burden making them impractical for real-world applications where speed or system resources were limited.
   - **RoBERTa and BERT:** Middle-ground models, RoBERTa being more efficient for nuanced text classification since its pretraining techniques are inattentive and more refined.

3. **Implications for Real-World Applications**

   These results have a number of ramifications for health care applications:

   - **Real-time Systems:** The efficiency of DistilBERT provides an opportunity for the model to succeed in clinical decision support systems, in which rapid predictions are critical.
   - **Contemporary Research:** They also tend to outperform BERT on Sentence Similarity or Classification, such as when the task involves clustering similar patients or comparing patients to standard measures.
   - **General-purpose Use:** For general healthcare NLP tasks, BERT continues to be a sound choice, achieving a nice compromise between performance and resource usage.

4. **Comparison with Existing Research**

   The results are consistent with previous studies highlighting the transformer models as state of the art for text classification in the healthcare domain. For example:

   - The superior recall and precision of RoBERTa seen in this study is in line with Peng et al. (2019)'s results that show RoBERTa is optimized for fine-grained text classification tasks.
   - This aligns with the general efficiency of DistilBERT as concluded by Sanh et al. (2020) and is indicative of DistilBERT's low resource nature.

   This study builds upon earlier work by filling a gap with a holistic comparison with four transformer models, providing pragmatic insights into their trade-offs towards healthcare.

5. **Limitations in Results**

   Although the findings are encouraging, some limitations were acknowledged:

- **Dataset-Specific Bias:** The tree-structured hierarchical nature and large textual contents of the PubMed dataset may have introduced bias for tree-structured models like XLNet and RoBERTa which cannot be generalized to other datasets
- **Computational Considerations:** The comparatively larger training time for XLNet has implications on its scalability in large scale healthcare.

6. **Future Directions**

   Based the observations above, future research may consider:

   - Increased applications of domain-specific pretraining approaches, such as BioBERT or ClinicalBERT tailored for healthcare datasets.
   - Who probably evaluating on this stuff, diversity electronic health records (EHRs) and patient-generated text improving generalizability.
   - Exploring hybrids that inherit computational efficiency from DistilBERT and contextual robustness from XLNet.

6.6.4 Broader Implications for NLP

This pick up of the discussion continues the research of transformer-based models for professional applications across different domains. These results also highlight the adaptability of these high performing models to hierarchical and multi-label classification, even in a very niche domain, such as health care. They also highlight the necessity for specific applications to select model(s) according to accuracy, computational efficiency, and resource availability.

6.6.5 Limitations and Future Directions

Thus, while making their study novel by making discovery for transformers on healthcare text classification, it had certain limitations:

- *Computational Constraints:* XLNet provides an implementation that takes a surprisingly long time and requires a lot of resources, and therefore it is necessary to develop more efficient computationally intensive models.
- *Dataset-Specific Evaluation:* The results reported will be specific to the PubMed dataset, and more general and in-depth understanding of the models capabilities can be gained by performing further evaluations on other healthcare-related datasets.

Future research directions:

- Incorporation of domain specific pretraining (i.e. BioBERT or ClinicalBERT) to further improve model performance.
- Evaluation of real world healthcare transformer applications, such as clinical note summarization or electronic health record (EHR) analysis.
- Integration of lightweight architectures (such as DistilBERT) with more complex models (such as XLNet) to enhance performance further.

7 Conclusion and Future Work

So, what transformation to expect from transformer models like BERT, RoBERTa, DistilBERT and XLNet on healthcare text classification, this study has demonstrated. This paper referenced real life health care applications, used the PubMed Multi-Label Text Classification Dataset to evaluate their performance, and analyzed the strengths and weaknesses of these models. DistilBERT was efficient and XLNet and RoBERTa were preferred when the number of contextual dependencies were complex but on multi-label classification layered models performed decently. Given that BERT is a foundational model, it actually represents a middle-ground model when it comes to performance and efficiency.

This study serves as a demonstration that transformer models not only improve performance across multiple classification of medical literature, summarization of clinical notes, and real-time patient query analyses. Yet these trade offs of computational efficiency for contextual fidelity do underscore the importance of deliberate model selection for any given application. So for example, the DistilBERT variants perform well in resource-constrained environments; on the other hand, XLNet could be far more effective in applications where a better contextual understanding is required.

What does somewhat temper these encouraging findings is a number of study limitations. This is very limited study on one representative dataset in absence of any foundational study to appropriately generalize any variation (known in real examples) across healthcare. The more recent use of the pre-trained transformer model without pre-training on domain-considerations of the domain of domain-tuned variants, such as BioBERT and ClinicalBERT, does not lead to optimization of the terminologies utilized in this area. A major limitation was the absence of comparisons with other models outside the realm of transformers that could lend insight into other model-specific advantages of the transformer architecture.

These predicates are some potential pathways for future research to broaden the analysis to include aim more categorical information, practically prepending clinical notes or patient-generated text that would crib the generalizability of the models. Domain-specific pretraining, e.g., weight-initialization on BioBERT or ClinicalBERT, would certainly yield better performance for such models on healthcare specific linguistic tasks. Similarly you should perform an ensemble algorithms investigation that is combining the advantages of this set of various transformer algorithms toward the generation of an aggregate algorithm that is useful by the majority of the objectives that require efficiency but a strong expertise behind it.

Future works can also explore interpretability and explainability which are really important features for deploying AI models in healthcare. The techniques described, such as attention heatmaps or layer-wise relevance propagation, might explain how models make decisions, and as such improve trust and acceptance among health professionals. The ability of transformer models to integrate into real-world health care workflows, whether through clinical decision support systems or automated patient record management, will provide a pragmatic avenue through which to validate the impact of transformer models on improvement in health care delivery.

In summary, this review has been able to deliver a picture of how transformer models can change the game for health care text classification as well as suggest directions for future work in both research and real-world settings. The future of AI in health care will be increasingly more rewarding when addressing its current shortcomings through

identified limitations that need to be mitigated in the patient care quality and medical practice.

References

Ahmed, S. S. and Madasamy, A. K. (2021). Classification of censored tweets in chinese language using xlnet, pp. 136–139.

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, Q., Naumann, T. and McDermott, M. (2019). Publicly available clinical bert embeddings, *arXiv preprint arXiv:1904.05342* .

Beltagy, I., Lo, K. and Cohan, A. (2019). Scibert: A pretrained language model for scientific text, *arXiv preprint arXiv:1903.10676* .

Dai, Z., Yang, Z., Le, Q. V. and Salakhutdinov, R. (2021). Xlnet and its applications in healthcare nlp: Challenges and opportunities, *Healthcare Informatics Journal* **10**(3): 15–29.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805* .

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019a). Bert: Pre-training of deep bidirectional transformers for language understanding.
URL: *https://arxiv.org/abs/1810.04805*

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019b). Bert: Pre-training of deep bidirectional transformers for language understanding.
URL: *https://arxiv.org/abs/1810.04805*

Hinton, G., Vinyals, O. and Dean, J. (2015). Distilling the knowledge in a neural network.
URL: *https://arxiv.org/abs/1503.02531*

Huang, K., Altosaar, J. and Ranganath, R. (2019a). Clinicalbert: Modeling clinical notes and predicting hospital readmission, *arXiv preprint arXiv:1904.05342* .

Huang, K., Altosaar, J. and Ranganath, R. (2019b). Clinicalbert: Modeling clinical notes and predicting hospital readmission, *arXiv preprint arXiv:1904.05342* .

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H. and Kang, J. (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining, *Bioinformatics* **36**(4): 1234–1240.

Liu, L., Zhang, M. and Wu, J. (2021). Evaluation of classification metrics for healthcare data, *Journal of Healthcare Informatics* **35**(4): 441–457.
URL: *https://doi.org/10.1016/j.jhi.2021.01.014*

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V. (2019a). Roberta: A robustly optimized bert pretraining approach, *arXiv preprint arXiv:1907.11692* .

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V. (2019b). Roberta: A robustly optimized bert pretraining approach. **URL:** *https://arxiv.org/abs/1907.11692*

Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J. and Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence, *Nature* **616**(7956): 259–265.

Nerella, S., Bandyopadhyay, S., Zhang, J., Contreras, M., Siegel, S., Bumin, A., Silva, B., Sena, J., Shickel, B., Bihorac, A., Khezeli, K. and Rashidi, P. (2024). Transformers and large language models in healthcare: A review, *Artificial Intelligence in Medicine* **154**: 102900.

Park, J. and Lee, H. (2023). Z-aligned3 dataset and applications in healthcare nlp, *AI in Medicine* .

Peng, Y., Yan, S. and Lu, Z. (2019). Transfer learning in biomedical natural language processing: An evaluation of bert and roberta, *arXiv preprint arXiv:1906.05474* .

Sanh, V., Debut, L., Chaumond, J. and Wolf, T. (2019). Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter, *arXiv preprint arXiv:1910.01108* .

Sanh, V., Debut, L., Chaumond, J. and Wolf, T. (2020). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.
URL: *https://arxiv.org/abs/1910.01108*

Si, J., Zhang, W. and Feng, L. (2019). Clinicalbert: Pretrained language model for clinical named entity recognition, *Computers in Biology and Medicine* **101**: 123–135.

Si, M., Zhao, W., Chen, Y. et al. (2019a). Evaluating interpretability in permutation-based nlp models: Case study of xlnet in healthcare, *Proceedings of the Annual Conference on Neural Information Processing Systems* **32**: 10–18.

Si, M., Zhao, W., Chen, Y. et al. (2019b). Evaluating interpretability in transformer models for high-stakes applications, *Proceedings of the Annual Conference on Neural Information Processing Systems* **32**: 10–18.

Sun, J., Wang, L. and Zhang, H. (2021). Transformers in document classification: Trade-offs and challenges, *Journal of Artificial Intelligence Research* **45**: 123–145.

Sun, Y., Li, J. and Wang, M. (2021). Comparative evaluation of transformer models for document classification in domain-specific applications, *Journal of Artificial Intelligence Research* **65**: 45–59.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. and Polosukhin, I. (2017). Attention is all you need, *Advances in Neural Information Processing Systems* **30**.

Yacouby, R. and Axman, D. (2020). Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models, *in* S. Eger, Y. Gao, M. Peyrard, W. Zhao and E. Hovy (eds), *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, Association for Computational Linguistics,

Online, pp. 79–91.
URL: *https://aclanthology.org/2020.eval4nlp-1.9*

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding, *Advances in Neural Information Processing Systems* **32**.

Zhao, M., Chen, X. and Liu, Y. (2020). Attention mechanisms in disease risk assessment: A comprehensive study, *International Journal of Medical Informatics* **87**: 56–72.

Zhao, S., Singh, A., Xu, K. et al. (2020a). Addressing the interpretability gap in transformer models for clinical applications, *Journal of Biomedical Informatics* **105**: 103421.

Zhao, S., Singh, A., Xu, K. et al. (2020b). Improving the utility of ehr data with xlnet for clinical sequence prediction tasks, *arXiv preprint arXiv:2005.12310* .